<Bioinformatics Center>Mathematical Bioinformatics

Bioinformatics Center
– Mathematical Bioinformatics –

http://www.bic.kyoto-u.ac.jp/takutsu/index.html

Prof
AKUTSU, Tatsuya
(D Eng)

Assist Prof
HAYASHIDA, Morihiro
(D Inf)

Assist Prof
TAMURA, Takeyuki
(D Inf)

Students

NAKAJIMA, Natsu (RF)	BAO, Yu (D2)	KAWAKAMI, Yuko (M1)
RUAN, Peiying (RF)	CAO, Yue (M2)	FUKUSAKO, Yuta (M1)
JIRA, Jindalertudomdee (D3)	NGOUV, Hayliang (M2)	LI, Ruiming (RS)

Guest Scholars

HOU, Wenpin (Ph D)	The University of Hong Kong, China, P.R., 25 March–22 August
MÜNZNER, Ulrike (Ph D)	Humboldt University, Germany, 15 April–15 June
CHENG, Xiaoqing (Ph D)	The University of Hong Kong, China, P.R., 22 June–13 July

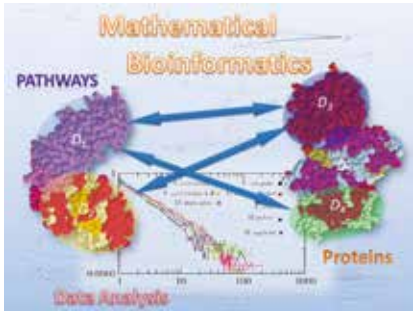
Guest Res Assoc

MELKMAN, Avraham (Ph D)	Ben Gurion University of the Negev, Israel, 18 February-18 April

Scope of Research

Due to rapid progress of genome sequencing technology, whole genome sequences of organisms ranging from bacteria to human have become available. In order to understand the meaning behind the genetic code, we have been developing algorithms and software tools for analyzing biological data based on advanced information technologies such as theory of algorithms, artificial intelligence, and machine learning. We are currently studying the following topics: systems biology, scale-free networks, protein structure prediction, inference of biological networks, chemo-informatics, and discrete and stochastic methods for bioinformatics.

KEYWORDS
Scale-free Networks
Boolean Networks
Chemical Graphs
Grammar-based Compression
Protein Complexes



Selected Publications

Akutsu, T.; Tamura, T.; Melkman, A. A.; Takasu, A., On the Complexity of Finding a Largest Common Subtree of Bounded Degree, *Theoretical Computer Science*, **590,** 2-16 (2015).

Nacher, J. C.; Akutsu, T., Structurally Robust Control of Complex Networks, *Physical Review E*, **91,** 012826 (2015).

Mori, T.; Takasu, A.; Jansson, J.; Hwang, J.; Tamura, T.; Akutsu, T., Similar Subtree Search Using Extended Tree Inclusion, *IEEE Transactions on Knowledge and Data Engineering*, **27,** 3360-3373 (2015).

Hayashida, M.; Jindalertudomdee, J.; Zhao, Y.; Akutsu, T., Parallelization of Enumerating Tree-like Chemical Compounds by Breadth-first Search Order, *BMC Medical Genomics*, **8(Suppl 2),** S15 (2015).

Kagami, H.; Akutsu, T.; Maegawa, S.; Hosokawa, S.; Nacher, J. C., Determining Associations between Human Diseases and non-coding RNAs with Critical Roles in Network Control, *Scienfitic Reports*, **5,** 14577 (2015).

Similar Subtree Search Using Extended Tree Inclusion

In this research, we consider the problem of identifying all locations of subtrees in a large tree or in a large collection of trees that are similar to a specified pattern tree, where all trees are assumed to be rooted and node-labeled. To calculate the similarity of two trees, tree edit distance is widely used, but it is NP-hard (Non-deterministic Polynomial-time hard) to compute for unordered trees. Therefore, we propose a new similarity measure that extends the concept of unordered tree inclusion by taking the costs of insertion and substitution operations on the pattern tree into account, and present an algorithm for computing it. The proposed algorithm has the same time complexity as the original one for unordered tree inclusion: i.e., it runs in $O(|T1||T2|)$ time, where $T1$ and $T2$ denote the pattern tree and the text tree, respectively, when the maximum outdegree of $T1$ is bounded by a constant. The experimental evaluation conducted using synthetic and real datasets confirms that the proposed algorithm is fast and scalable and very useful for bibliographic matching, which is a typical entity resolution problem for tree-structured data. Moreover, we extend our algorithm to also allow a constant number of deletion operations on $T1$ while still running in $O(|T1||T2|)$ time.

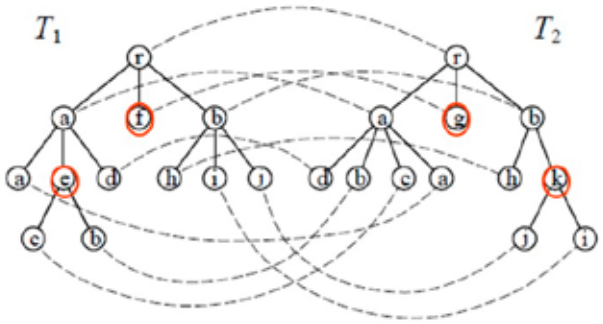


Figure 1. In unordered trees, T_2 is obtained by deletion of the node 'e', insertion of a node 'k', and substitutions of nodes 'f' to 'g' from T_1.

Parallelization of Enumerating Tree-like Chemical Compounds by Breadth-first Search Order

Enumeration of chemical structures is useful for designing and finding new drugs, and determining chemical structures from mass spectrometry. We previously developed efficient algorithms, BfsMulEnum and BfsSimEnum, for enumerating tree-like chemical compounds with and without multiple bonds, respectively. In many instances, the algorithms were able to enumerate chemical structures faster than other existing methods.

Modern computers have multiple processing cores, and are able to execute many tasks simultaneously. In this work, we developed three parallelized algorithms, BfsEnumP1–3, by modifying BfsSimEnum to further reduce execution time. BfsSimEnum constructs a family tree in which each vertex denotes a molecular tree. BfsEnumP1–3 divide a set of vertices with some given depth of the family tree into several subsets, each of which is assigned to a core.

For evaluation, we performed several experiments with varying division depth and number of cores, and showed that BfsEnumP1–3 are useful to reduce the execution time for enumeration of tree-like chemical compounds. In addition, we showed that BfsEnumP3 achieved more than 80% parallelization efficiency using up to 11 cores, and reduced the execution time using 12 cores to about 1/10 of that by BfsSimEnum.

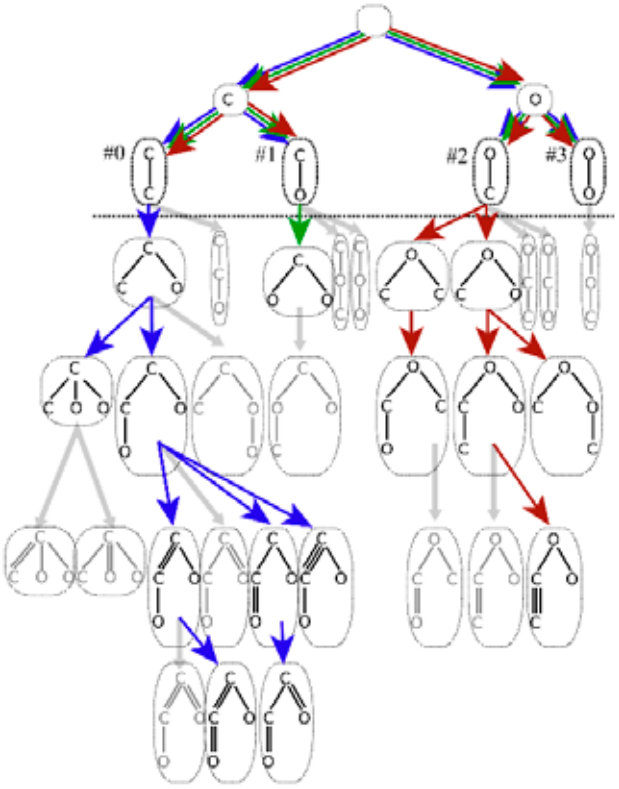


Figure 2. Example of a family tree created by BfsSimEnum and BfsMulEnum for $C_2O_2H_2$ and its separation by BfsEnumP1 with three cores and division depth 2.